Robust Outlier Rejection for 3D Registration with Variational Bayes

Haobo Jiang[1], Zheng Dang[2], Zhen Wei[2], Jin Xie[*1], Jian Yang[*1], and Mathieu Salzmann[*2]

[1]PCA Lab, Nanjing University of Science and Technology, China
[2]CVLab, EPFL, Switzerland
{jiang.hao.bo, csjxie, csjyang}@njust.edu.cn
{zheng.dang, zhen.wei, mathieu.salzmann}@epfl.ch

Abstract

Learning-based outlier (mismatched correspondence) rejection for robust 3D registration generally formulates the outlier removal as an inlier/outlier classification problem. The core for this to be successful is to learn the discriminative inlier/outlier feature representations. In this paper, we develop a novel variational non-local network-based outlier rejection framework for robust alignment. By reformulating the non-local feature learning with variational Bayesian inference, the Bayesian-driven long-range dependencies can be modeled to aggregate discriminative geometric context information for inlier/outlier distinction. Specifically, to achieve such Bayesian-driven contextual dependencies, each query/key/value component in our non-local network predicts a prior feature distribution and a posterior one. Embedded with the inlier/outlier label, the posterior feature distribution is label-dependent and discriminative. Thus, pushing the prior to be close to the discriminative posterior in the training step enables the features sampled from this prior at test time to model high-quality long-range dependencies. Notably, to achieve effective posterior feature guidance, a specific probabilistic graphical model is designed over our non-local model, which lets us derive a variational low bound as our optimization objective for model training. Finally, we propose a voting-based inlier searching strategy to cluster the high-quality hypothetical inliers for transformation estimation. Extensive experiments on 3DMatch, 3DLoMatch, and KITTI datasets verify the effectiveness of our method. Code is available at https://github.com/Jiang-HB/VBReg.

---
*Corresponding authors

Haobo Jiang, Jin Xie, and Jian Yang are with PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, China.

1. Introduction

Point cloud registration is a fundamental but challenging 3D computer vision task, with many potential applications such as 3D scene reconstruction [1, 39], object pose estimation [13, 44], and Lidar SLAM [15, 51]. It aims to align two partially overlapping point clouds by estimating their relative rigid transformation, i.e., 3D rotation and 3D translation. A popular approach to address the large-scale scene registration problem consists of extracting point descriptors [11, 14, 17, 37, 38, 50] and establishing correspondences between the two point clouds, from which the transformation can be obtained geometrically. In this context, much effort has been dedicated to designing traditional and deep learning-based descriptors [3, 11, 21, 41, 50]. However, the resulting correspondences inevitably still suffer from outliers (wrong matchings), particularly in challenging cases, such as low-overlap, repetitive structures, or noisy point sets, leading to registration failure.

To address this, many outlier filtering strategies have been developed to robustify the registration process. These include traditional rejection methods using random sample consensus [16], point-wise descriptor similarity [7, 32] or group-wise spatial consistency [46]. Deep learning methods have also been proposed, focusing on learning correspondence features used to estimate inlier confidence values [2, 10, 33]. In particular, the current state-of-the-art method, PointDSC [2], relies on a spatial consistency-driven non-local network to capture long-range context in its learned correspondence features. While effective, PointDSC still yields limited registration robustness, particularly for scenes with a high outlier ratio, where the spatial consistency constraints may become ambiguous [36], thereby degrading the correspondence features' quality.

In this paper, we propose to explicitly account for the ambiguities arising from high outlier ratios by developing a probabilistic feature learning framework. To this end, we introduce a variational non-local network based on an at-

tention mechanism to learn discriminative inlier/outlier feature representations for robust outlier rejection. Specifically, to capture the ambiguous nature of long-range contextual dependencies, we inject a random feature in each query, key, and value component in our non-local network. The prior/posterior distributions of such random features are predicted by prior/posterior encoders. To encourage the resulting features to be discriminative, we make the posterior feature distribution label-dependent. During training, we then push the prior distribution close to the label-dependent posterior, thus allowing the prior encoder to also learn discriminative query, key, and value features. This enables the features sampled from this prior at test time to model high-quality long-range dependencies.

To achieve effective variational inference, we customize a probabilistic graphical model over our variational non-local network to characterize the conditional dependencies of the random features. This lets us derive a variational lower bound as the optimization objective for network training. Finally, we propose a voting-based deterministic inlier searching mechanism for transformation estimation, where the correspondence features learned from all non-local iterations jointly vote for high-confidence *hypothetical inliers* for SVD-based transformation estimation. We theoretically analyze the robustness of our deterministic inlier searching strategy compared to RANSAC, which also motivates us to design a conservative seed selection mechanism to improve robustness in sparse point clouds.

To summarize, our contributions are as follows:

- We propose a novel variational non-local network for outlier rejection, learning discriminative correspondence features with Bayesian-driven long-range contextual dependencies.

- We customize the probabilistic graphical model over our variational non-local network and derive the variational low bound for effective model optimization.

- We introduce a Wilson score-based voting mechanism to search high-quality *hypothetical inliers*, and theoretically demonstrate its superiority over RANSAC.

Our experimental results on extensive benchmark datasets demonstrate that our framework outperforms the state-of-the-art registration methods.

2. Related Work

End-to-end Registration Methods. With the advances of deep learning in the 3D vision field [34], the learning-based end-to-end registration model has achieved increasing research attention. DCP [42] uses the feature similarity to establish pseudo correspondences for SVD-based transformation estimation. RPM-Net [47] exploits the Sinkhorn

layer and annealing for discriminative matching map generation. [22, 23] integrate the cross-entropy method into the deep model for robust registration. RIENet [40] uses the structure difference between the source neighborhood and the pseudo-target one for inlier confidence evaluation. With the powerful feature representation of Transformer, RegTR [48] effectively aligns large-scale indoor scenes in an end-to-end manner. [13] propose a match-normalization layer for robust registration in the real-world 6D object pose estimation task. More end-to-end models such as [10, 18, 28–30, 33, 53] also present impressive precisions.

Learning-based Feature Descriptors. To align the complex scenes, a popular pipeline is to exploit feature descriptors for 3D matching. Compared to hand-crafted descriptors such as [17, 37, 38], the deep feature descriptor presents superior registration precision and has achieved much more attention in recent years. The pioneering 3DMatch [50] exploits the Siamese 3D CNN to learn the local geometric feature via contrastive loss. FCGF [11] exploits a fully convolutional network for dense feature extraction in a one-shot fashion. Furthermore, D3feat [3] jointly learns the dense feature descriptor and the detection score for each point. By integrating the overlap-attention module into D3feat, Predator [21] largely improves the registration reliability in low-overlapping point clouds. YOHO [41] utilizes the group equivariant feature learning to achieve the rotation invariance and shows great robustness to the point density and the noise interference. [35] develops a geometric transformer to learn the geometric context for robust super-point matching. Lepard [31] embeds the relative 3D positional encoding into the transformer for discriminative descriptor learning.

Outlier Rejection Methods. Despite significant progress in learning-based feature descriptor, generating mismatched correspondences (outliers) in some challenging scenes remains unavoidable. Traditional outlier filtering methods, such as RANSAC [16] and its variants [4, 24, 27], use repeated sampling and verification for outlier rejection. However, these methods tend to have a high time cost, particularly in scenes with a high outlier ratio. Instead, FGR [52] and TEASER [45] integrate the robust loss function into the optimization objective to weaken the interference from outliers. Recently, Chen et al. [9] developed second-order spatial compatibility for robust consensus sampling. With the rise of deep 3D vision, most learnable outlier rejection models [10,33] formulate outlier rejection as a binary classification task and reject correspondences with low confidence. Yi et al. [49] proposed a context normalization-embedded deep network for inlier evaluation, while Brachmann et al. [6] enhanced classical RANSAC with neural-guided prior confidence. As our baseline, PointDSC [2] proposes exploiting a spatial consistency-guided non-local inlier classifier for inlier evaluation, followed by neural spectral matching for robust registration. However, under high outlier

ratios, spatial consistency can be ambiguous (as shown in Fig. 1), misleading non-local feature aggregation. Instead, we propose exploiting Bayesian-driven long-range dependencies for discriminative non-local feature learning.

3. Approach

3.1. Background

Problem Setting. In the pairwise 3D registration task, given a source point cloud $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^3 \mid i = 1, ..., |\mathbf{X}|\}$ and a target point cloud $\mathbf{Y} = \{\mathbf{y}_j \in \mathbb{R}^3 \mid j = 1, ..., |\mathbf{Y}|\}$, we aim to find their optimal rigid transformation consisting of a rotation matrix $\mathbf{R}^* \in SO(3)$ and a translation vector $\mathbf{t}^* \in \mathbb{R}^3$ to align their overlapping region precisely. In this work, we focus on the descriptor-based pipeline for large-scale scene registration. Based on the feature-level nearest neighbor, we construct a set of putative correspondences $\mathcal{C} = \{\mathbf{c}_i = (\mathbf{x}_i, \mathbf{y}_i) \in \mathbb{R}^6 \mid i = 1, ..., |\mathcal{C}|\}$. The inlier (correctly matched correspondence) is defined as the correspondence satisfying $\|\mathbf{R}^*\mathbf{x}_i + \mathbf{t}^* - \mathbf{y}_i\| < \varepsilon$, where ε indicates the inlier threshold.

Vanilla Non-local Feature Embedding. Given the putative correspondence set $\mathcal{C}$, [2] leverages the spatial consistency-guided non-local network (*SCNonlocal*) for their feature embedding. The injected geometric compatibility matrix can effectively regularize the long-range dependencies for discriminative inlier/outlier feature learning. In detail, it contains L iterations and the feature aggregation in l-th iteration can be formulated as:

$$\mathbf{F}_i^{l+1} = \mathbf{F}_i^l + \text{MLP}\left(\sum_{j=1}^{|\mathcal{C}|} \text{softmax}_j(\boldsymbol{\alpha}^l \boldsymbol{\beta}) \mathbf{V}_j^l\right), \quad (1)$$

where $\mathbf{F}_i^l \in \mathbb{R}^d$ indicates the feature embedding of correspondence $\mathbf{c}_i$ in l-th iteration (the initial feature $\mathbf{F}_i^0$ is obtained via linear projection on $\mathbf{c}_i$) and $\mathbf{V}_i^l = f_v^l(\mathbf{F}_i^l) \in \mathbb{R}^d$ is the projected value feature. $\boldsymbol{\alpha}^l \in \mathbb{R}^{|\mathcal{C}| \times |\mathcal{C}|}$ is the non-local attention map whose entry $\boldsymbol{\alpha}_{i,j}^l$ reflects the feature similarity between the projected query feature $\mathbf{Q}_i^l = f_q^l(\mathbf{F}_i^l) \in \mathbb{R}^d$ and the key feature $\mathbf{K}_i^l = f_k^l(\mathbf{F}_j^l) \in \mathbb{R}^d$. $\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{C}| \times |\mathcal{C}|}$ represents the geometric compatibility matrix of correspondences, where the compatibility between $\mathbf{c}_i$ and $\mathbf{c}_j$ is:

$$\boldsymbol{\beta}_{i,j} = \max\left(0, 1 - \frac{d_{ij}^2}{\varepsilon^2}\right), \ d_{ij} = |\|\mathbf{x}_i - \mathbf{x}_j\| - \|\mathbf{y}_i - \mathbf{y}_j\||. \quad (2)$$

Based on the fact that the geometric distance $d_{i,j}$ of inliers $\mathbf{c}_i$ and $\mathbf{c}_j$ tend to be minor, Eq. 2 will assign a high compatibility value on the inlier pair, thereby promoting the non-local network to effectively cluster the inlier features for discriminative inlier/outlier feature learning.

3.2. Variational Non-local Feature Embedding

While effective, *SCNonlocal* still suffers from ambiguous long-range dependencies, especially in some challeng-
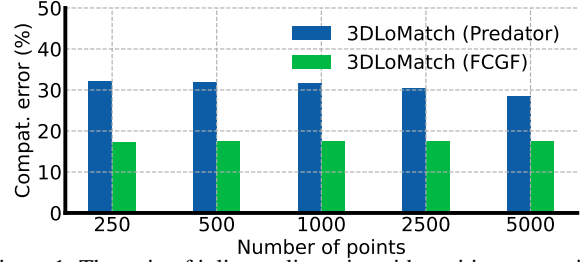

Figure 1. The ratio of inlier-outlier pairs with positive compatibilities in 3DLoMatch [21] using FCGF and Predator descriptors.

ing scenes (e.g., the low-overlapping case). Two essential reasons are: **(i)** Wrong geometric compatibility. As shown in Fig. 1, for 3DLoMatch dataset with Predator and FCGF descriptors, almost 30% and 17% of inlier-outlier pairs own the positive compatibility values, respectively, which potentially misleads the attention weight for wrong feature clustering. **(ii)** Lack of uncertainty modeling. In symmetric or repetitive scenes, the inlier/outlier prediction contains significant uncertainty. Therefore, it's necessary to design a robust feature representation to capture such uncertainty.

To overcome them, we develop a variational non-local network, a probabilistic feature learning framework, for discriminative correspondence embedding. Our core idea is to inject random features into our model to capture the ambiguous nature of long-range dependencies, and then leverage the variational Bayesian inference to model the Bayesian-driven long-range dependencies for discriminative feature aggregation. Specifically, we first introduce the random feature variables $z_{k,i}^l$, $z_{q,i}^l$ and $z_{v,i}^l$ into the key $\mathbf{K}_i^l$, query $\mathbf{Q}_i^l$ and value $\mathbf{V}_i^l$ components in our non-local module to capture their potential uncertainty in the long-range dependency modeling. Then, the prior/posterior encoders are constructed to predict their prior feature distribution and the posterior one, respectively. Embedded with the inlier/outlier label, the posterior feature distribution is label-dependent and discriminative. Thus, by pushing the prior close to the discriminative posterior in the training phase, this prior at test time also tends to sample discriminative query, key, and value features for high-quality long-range dependency modeling.

Probabilistic Graphical Model over Variational Non-local Network. To achieve effective variational Bayesian inference, we need to first characterize the conditional dependencies of the injected random features so that the variational lower bound can be derived as the optimization objective for model training. As shown in Fig. 2, we customize the probabilistic graphical model over our non-local network to clarify the dependencies of random features (the circles). The solid line denotes the label prediction process, while the dashed line represents our label-dependent posterior encoder (*i.e.*, inference model). Notably, the deterministic hidden query/key/value features $\mathbf{h}_{k,i}^l \in \mathbb{R}^{d'}$,

$\mathbf{h}_{q,i}^{l} \in \mathbb{R}^{d'}$, and $\mathbf{h}_{v,i}^{l} \in \mathbb{R}^{d'}$ are also introduced to summarize the historical information for better feature updating in each iteration.

Inlier/outlier Prediction Process. Based on the defined conditional dependencies in Fig. 2, the prediction process of correspondence labels $\mathbf{b} = \{b_1, b_2, ..., b_{|\mathcal{C}|} \mid b_i \in \{0, 1\}\}$ (1 indicates inlier and 0 outlier) is formulated as follows. Beginning with the initial linear projection $\tilde{\mathbf{F}}^0 \in \mathbb{R}^{|\mathcal{C}| \times d}$ of correspondences $\mathcal{C} = \{\mathbf{c}_i\}$, we iteratively perform the probabilistic non-local aggregation for feature updating. In the l-th iteration, we first employ a Gated Recurrent Unit (GRU) [12] to predict the hidden query/key/value features which summarize the historical query/key/value features (sampled from the prior distributions) and the correspondence features in previous iterations:

$$
\begin{aligned}
\mathbf{h}_{q,i}^{l} &= \mathrm{GRU}_q(\mathbf{h}_{q,i}^{l-1}, [\mathbf{z}_{q,i}^{l-1}, \tilde{\mathbf{F}}_i^{l-1}]), \\
\mathbf{h}_{k,i}^{l} &= \mathrm{GRU}_k(\mathbf{h}_{k,i}^{l-1}, [\mathbf{z}_{k,i}^{l-1}, \tilde{\mathbf{F}}_i^{l-1}]), \\
\mathbf{h}_{v,i}^{l} &= \mathrm{GRU}_v(\mathbf{h}_{v,i}^{l-1}, [\mathbf{z}_{v,i}^{l-1}, \tilde{\mathbf{F}}_i^{l-1}]),
\end{aligned}
\tag{3}
$$

where $[\cdot, \cdot]$ denotes the feature concatenation and $\tilde{\mathbf{F}}_i^{l-1}$ is the learned correspondence features of $\mathbf{c}_i$ in iteration $l-1$. Then, with as input the predicted hidden features, the prior encoder $p_\theta(\cdot)$ is utilized to predict the prior feature distribution for query/key/value, respectively. Furthermore, we sample features $\mathbf{z}_{q,i}^{l} \in \mathbb{R}^{\tilde{d}}$, $\mathbf{z}_{k,i}^{l} \in \mathbb{R}^{\tilde{d}}$ and $\mathbf{z}_{v,i}^{l} \in \mathbb{R}^{\tilde{d}}$ from the predicted prior query/key/value distribution and combine them with the hidden features to predict the corresponding query $\tilde{\mathbf{Q}}_i^{l} \in \mathbb{R}^d$, key $\tilde{\mathbf{K}}_i^{l} \in \mathbb{R}^d$ and value $\tilde{\mathbf{V}}_i^{l} \in \mathbb{R}^d$ through a neural network $f_\theta^{q,k,v} : \mathbb{R}^{d'+\tilde{d}} \to \mathbb{R}^d$:

$$
\begin{aligned}
\mathbf{z}_{q,i}^{l} &\sim p_\theta(z_{q,i}^{l} \mid \mathbf{h}_{q,i}^{l}), & \tilde{\mathbf{Q}}_i^{l} &= f_\theta^{q}([\mathbf{z}_{q,i}^{l}, \mathbf{h}_{q,i}^{l}]), \\
\mathbf{z}_{k,i}^{l} &\sim p_\theta(z_{k,i}^{l} \mid \mathbf{h}_{k,i}^{l}), & \tilde{\mathbf{K}}_i^{l} &= f_\theta^{k}([\mathbf{z}_{k,i}^{l}, \mathbf{h}_{k,i}^{l}]), \\
\mathbf{z}_{v,i}^{l} &\sim p_\theta(z_{v,i}^{l} \mid \mathbf{h}_{v,i}^{l}), & \tilde{\mathbf{V}}_i^{l} &= f_\theta^{v}([\mathbf{z}_{v,i}^{l}, \mathbf{h}_{v,i}^{l}]),
\end{aligned}
\tag{4}
$$

where the prior feature distribution is the Gaussian distribution with the mean and the standard deviation parameterized by a neural network. Finally, with the learned $\tilde{\mathbf{Q}}_i^{l}$, $\tilde{\mathbf{K}}_i^{l}$ and $\tilde{\mathbf{V}}_i^{l}$, the correspondence feature $\tilde{\mathbf{F}}_i^{l}$ in l-th iteration can be aggregated with the same non-local operation in Eq. 1. After L feature iterations, we feed the correspondence feature $\tilde{\mathbf{F}}_i^{L}$ in the last iteration into a label prediction model y_θ to predict the inlier/outlier labels $b_i \sim y_\theta(b_i \mid \tilde{\mathbf{F}}_i^{L})$, where the label prediction model outputs a scalar Gaussian distribution with the mean parameterized by the neural network and the unit variance.

Variational Posterior Encoder. Due to the nonlinearity of our variational non-local model, we cannot directly derive the precise posterior distribution for random query/key/value features using the standard Bayes' theorem. Taking inspiration from the Variational Bayesian inference, we construct a label-dependent posterior encoder
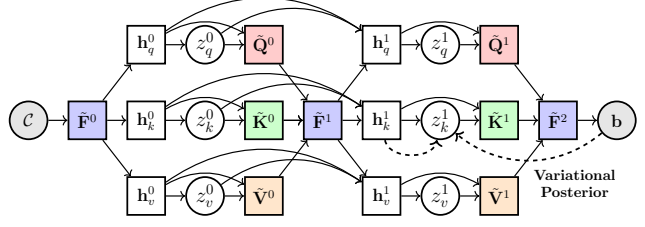

Figure 2. Probabilistic graphical model for our variational non-local network. For simplicity, we just demonstrate two iterations. The white circles indicate the random features and the white squares denote the deterministic hidden features. The solid line represents the inlier/outlier prediction process and the dashed line denotes the label-dependent variational posterior encoder. We just show the variational posterior for z_k^1.

$q_\phi(\cdot)$ to to approximate the feature posterior:

$$
\begin{aligned}
\mathbf{z}_{q,i}^{l} &\sim q_\phi(z_{q,i}^{l} \mid [\mathbf{h}_{q,i}^{l}, [b_i]_{\times k}]) \\
\mathbf{z}_{k,i}^{l} &\sim q_\phi(z_{k,i}^{l} \mid [\mathbf{h}_{k,i}^{l}, [b_i]_{\times k}]) \\
\mathbf{z}_{v,i}^{l} &\sim q_\phi(z_{v,i}^{l} \mid [\mathbf{h}_{v,i}^{l}, [b_i]_{\times k}]),
\end{aligned}
\tag{5}
$$

where $[b_i]_{\times k}$ indicates a label vector generated by tiling the scalar label k times. The output of each posterior encoder is a diagonal Gaussian distribution with parameterized mean and standard deviation.

Variational Lower Bound. Finally, we derive the optimization objective $\mathrm{ELBO}(\theta, \phi)$, the variational (evidence) lower bound of log-likelihood correspondence labels $\ln y_\theta(\mathbf{b} \mid \mathcal{C})$, to train our variational non-local network (Please refer to Appendix A for the detailed derivation):

$$
\begin{aligned}
\mathrm{ELBO}(\theta, \phi) = \mathbb{E}_{\prod_{l=0}^{L-1} q_\phi(z_{q,k,v}^{l} \mid \mathbf{h}_{q,k,v}^{l}, \mathbf{b})} & \left[\ln y_\theta(\mathbf{b} \mid \tilde{\mathbf{F}}^L) \right] - \\
\sum_{l=0}^{L-1} \mathbb{E}_{q_\phi} \left[\mathrm{D}_{\mathrm{KL}} \left(q_\phi(z_{q,k,v}^{l} \mid \mathbf{h}_{q,k,v}^{l}, \mathbf{b}) \| p_\theta(z_{q,k,v}^{l} \mid \mathbf{h}_{q,k,v}^{l}) \right) \right] &
\end{aligned}
\tag{6}
$$

where for clarity, we utilize the subscript q, k, v to denote the same operator performed on query/key/value. $\mathrm{D}_{\mathrm{KL}}(\cdot \| \cdot)$ denotes the Kullback–Leibler (KL) divergence between two distributions. By maximizing the variational lower bound above, we can optimize the network parameters to indirectly maximize the log-likelihood value of correspondence labels. Eq. 6 indicates that the discriminative, label-dependent feature posterior explicitly constrains the prior by reducing their KL divergence in the training phase, which promotes the query, key, and value features sampled from the prior to model the high-quality long-term dependencies at test time.

3.3. Voting-based Inlier Searching

With the learned correspondence features above, we then propose a voting-based sampling strategy to search the desired inlier subset from the entire putative correspondences

for optimal transformation estimation. Our sampling mechanism is deterministic and efficient. We first select a set of high-confidence seeds $\mathcal{C}_{seed}$ from $\mathcal{C}$ based on their inlier confidence (predicted in § 3.2) and the Non-Maximum Suppression (as performed in [2]), where the number of seeds $|\mathcal{C}_{seed}| = \lfloor |C| * v \rfloor$ (v is a fixed seed ratio). Then, for each seed, we cluster its most compatible correspondences into it to form the *hypothetical inliers*. Ideally, if the seed is an inlier and the compatible value is correct (without ambiguity problem as in Fig. 2), its most compatible correspondences are also inliers, thus we can cluster the desired inlier subset successfully. To cluster high-confidence *hypothetical inliers*, we perform a coarse-to-fine clustering strategy, including the voting-based coarse clustering and the Wilson score-based fine clustering.

Voting-based Coarse Clustering. We view the correspondence features $\{\tilde{\mathbf{F}}^1, ..., \tilde{\mathbf{F}}^L\}$ learned from all non-local iterations as L voters. For l-th voter, it deterministically clusters $\kappa - 1$ most compatible correspondences $^{(\kappa-1)}\mathcal{A}_{\mathbf{c}_i}^l$ for each seed $\mathbf{c}_i \in \mathcal{C}_{seed}$ to from the *hypothetical inliers* $^{(\kappa)}\mathcal{M}_{\mathbf{c}_i}^l = \{\mathbf{c}_i\} \cup {}^{(\kappa-1)}\mathcal{A}_{\mathbf{c}_i}^l$. To this end, we first compute the compatibility matrix $\mathbf{S}^l \in \mathbb{R}^{|\mathcal{C}| \times |\mathcal{C}|}$ for l-th voter, where each entry $\mathbf{S}_{i,j}^l = \text{clip}(1 - (1 - cos(\tilde{\mathbf{F}}_i^l, \tilde{\mathbf{F}}_j^l)/\sigma^2), 0, 1) \cdot \boldsymbol{\beta}_{i,j}$, where the parameter σ is to control the feature compatibility and geometric compatibility $\boldsymbol{\beta}_{i,j}$ is defined in Eq. 2. Thus, $^{(\kappa-1)}\mathcal{A}_{\mathbf{c}_i}^l$ can be determined as $\{\mathbf{c}_j \mid \mathbf{S}_{i,j}^l \geq \gamma_{\kappa-1}\}$, where threshold $\gamma_{\kappa-1}$ is the $(\kappa-1)$-th largest compatibility. Benefitting from our discriminative feature representations, $\mathbf{S}^l$ can effectively handle the ambiguity problem in Fig. 2 and promote robust inlier clustering. After the voting process above, each seed $\mathbf{c}_i$ can achieve L candidate *hypothetical inliers* $\{^{(\kappa)}\mathcal{M}_{\mathbf{c}_i}^1, ..., {}^{(\kappa)}\mathcal{M}_{\mathbf{c}_i}^L\}$.

Wilson score-based Fine Clustering. Based on the voted *hypothetical inliers* above, we then fuse these voted results with Wilson score [43] to form the high-quality *hypothetical inliers* $^{(\kappa)}\tilde{\mathcal{M}}_{\mathbf{c}_i}$. In detail. we denote $\mathcal{I}_{i,j}^l = \mathbb{I}\{\mathbf{c}_j \in {}^{(\kappa)}\mathcal{M}_{\mathbf{c}_i}^l\}$ to indicate whether $^{(\kappa)}\mathcal{M}_{\mathbf{c}_i}^l$ contains $\mathbf{c}_j$. The Wilson score $\mathbf{W}_{i,j}^n$ of accepting $\mathbf{c}_j$ into $^{(\kappa)}\tilde{\mathcal{M}}_{\mathbf{c}_i}$ can be computed as:

$$\mathbf{W}_{i,j}^n = \frac{1}{1 + \frac{z^2}{n}} \left[\hat{p}_{i,j}^{(n)} + \frac{z^2}{2n} - z\sqrt{\frac{\hat{p}_{i,j}^{(n)}(1 - \hat{p}_{i,j}^{(n)})}{n} + \frac{z^2}{4n^2}} \right], \tag{7}$$

where $\hat{p}_{i,j}^{(n)} = \frac{1}{n}\sum_{\tau=1}^{n} \mathcal{I}_{i,j}^\tau$ is the average acceptation ratio of top n voters and $z = 1.96$ (*i.e.*, z-score at 95% confidence level). Eq. 7 indicates that the Wilson score not only considers the sample mean but also the confidence (positively related to sample number n). Finally, among the set of Wilson scores under different sample numbers $\{\mathbf{W}_{i,j}^1, ..., \mathbf{W}_{i,j}^L\}$, we choose the largest one as the final Wilson score $\tilde{\mathbf{W}}_{i,j} = \max_{n \in \{1,...,L\}} \mathbf{W}_{i,j}^n$ for $\mathbf{c}_j$. Thus, the final *hypothetical inliers* set of $\mathbf{c}_i$ can be determined as

$^{(\kappa)}\tilde{\mathcal{M}}_{\mathbf{c}_i} = \{\mathbf{c}_i\} \cup \{\mathbf{c}_j \mid \tilde{\mathbf{W}}_{i,j} \geq \gamma'_{\kappa-1}\}$, where $\gamma'_{\kappa-1}$ is the $(\kappa - 1)$-th largest Wilson score.

Theoretical Analysis and Conservative Seed Selection. Finally, we try to theoretically analyze our deterministic inlier searching mechanism compared to RANSAC [16] and further propose a simple but effective conservative seed selection strategy for more robust 3D registration in sparse point clouds. We let $\{^{(\kappa)}\mathcal{M}_i^{sac}\}_{i=1}^J$ be the randomly sampled *hypothetical inlier* subset in RANSAC, and let $\mathcal{C}_{in}$, $\mathcal{C}_{out}$ and p_{in} be the inlier subset, outlier subset and the inlier ratio, respectively, $\mathcal{C} = \mathcal{C}_{in} \cup \mathcal{C}_{out}$, $p_{in} = |\mathcal{C}_{in}|/|\mathcal{C}|$. We also denote the inliers in seeds as $\tilde{\mathcal{C}}_{in} = \mathcal{C}_{in} \cap \mathcal{C}_{seed}$. Then, we can derive the following theorem (Please refer to Appendix B for the derivation process).

Theorem 1 *Assume the number of outliers in* $^{(\kappa)}\tilde{\mathcal{M}}_{\mathbf{c}_i}$ ($\mathbf{c}_i \in \tilde{\mathcal{C}}_{in}$) *follows a Poisson distribution* $Pois(\alpha \cdot \kappa)$. *Then, if* $\alpha < -\frac{1}{\kappa} \cdot \log\left[1 - (1 - p_{in}^\kappa)^{J/|\tilde{\mathcal{C}}_{in}|}\right] \triangleq \mathcal{U}$, *the probability of our method achieving the inlier subset is greater than or equal to that of RANSAC.*

$$P\left(\max_{\mathbf{c}_i \in \mathcal{C}_{seed}} |^{(\kappa)}\tilde{\mathcal{M}}_{\mathbf{c}_i} \cap \mathcal{C}_{in}| = \kappa\right) \geq$$
$$P\left(\max_{1 \leq i \leq J} |^{(\kappa)}\mathcal{M}_i^{sac} \cap \mathcal{C}_{in}| = \kappa\right). \tag{8}$$

Theorem 1 shows that as the inlier ratio p_{in} degrades, the upper bound $\mathcal{U}$ will increase, indicating our method tends to be more likely to achieve the inlier subset than RANSAC in high outlier cases. Therefore, our method tends to be more robust to outliers than RANSAC. In addition, Theorem 1 also shows that the more inliers $|\tilde{\mathcal{C}}_{in}|$ in $\mathcal{C}_{seed}$, the better for our method. It means selecting all correspondences in $\mathcal{C}$ as seeds theoretically seems the best choice since it can avoid missing any inlier. However, in real implementation, we have to perform inlier selection to accelerate the registration speed. Nevertheless, in sparse point clouds, the number of correspondences is small. If we still perform the original inlier selection, fewer seeds even may not contain any inlier, leading to registration failure. To overcome it, we develop a conservative seed selection strategy, which changes the seed number to $\max\{\lfloor v \cdot |\mathcal{C}| \rfloor, n\}$ ($n < |\mathcal{C}|$), where the lower bar n of the seed number can effectively avoid selecting too fewer inliers in sparse cases and in dense cases, it would degrade to the original selection strategy.

3.4. Rigid Transformation Estimation

With the voted *hypothetical inliers* $\{^{(\kappa)}\tilde{\mathcal{M}}_{\mathbf{c}_i} \mid \mathbf{c}_i \in \mathcal{C}_{seed}\}$ above, we estimate the optimal transformation parameter based on the *Procrustes method* [20] to minimize the least-squares errors for each *hypothetical inlier* group:

$$\hat{\mathbf{R}}_i, \hat{\mathbf{t}}_i = \arg\min_{\mathbf{R}, \mathbf{t}} \sum_{\mathbf{c}_j \in {}^{(\kappa)}\tilde{\mathcal{M}}_{\mathbf{c}_i}} \omega_j \cdot \|\mathbf{R}^\top \mathbf{x}_j + \mathbf{t} - \mathbf{y}_j\|_2, \tag{9}$$

where ω_j is the error weight computed by the neural spectral matching as in [2]. Then, we select the transformation

parameter that maximizes the number of overlapped correspondences as the final optimal transformation estimation:

$$\hat{\mathbf{R}}^*, \hat{\mathbf{t}}^* = \underset{\{\hat{\mathbf{R}}_i, \hat{\mathbf{t}}_i\}_{i=1}^{|\mathcal{C}|}}{\arg\max} \sum_{\mathbf{c}_j \in \mathcal{C}} v_i \cdot \mathbb{I}\left\{\|\hat{\mathbf{R}}_i^\top \mathbf{x}_j + \hat{\mathbf{t}}_i - \mathbf{y}_j\|_2 < \varepsilon\right\},$$

(10)

where $v_i = 1 - \|\hat{\mathbf{R}}_i^\top \mathbf{x}_j + \hat{\mathbf{t}}_i - \mathbf{y}_j\|_2^2/\varepsilon^2$ is used to re-weight the inlier count as performed in [23]. Finally, we refine it using all recovered inliers in a least-squares optimization as a common practice in [2, 4].

4. Experiments

4.1. Experimental Settings

Implementation Details. For our variational non-local module, the number of iterations L is 12, and the dimensions of the correspondence feature, random feature, and hidden feature are set to 128, 128, and 256, respectively. For our voting-based inlier sampling module, the size of *hypothetical inliers* κ is 40. For seed selection, seed ratio v is 0.1 and the lower bar of seed number n is 1000. Our model is trained with 50 epochs using Adam optimizer with learning rate 10^{-4} and weight decay 10^{-6}. We utilize PyTorch to implement our project and perform all experiments on the server equipped with an Intel i5 2.2 GHz CPU and one Tesla V100 GPU. For simplicity, we name our **V**ariational **B**ayesian-based **Reg**istration framework as **VBReg**.

Evaluation Metric. We use three metrics to evaluate our method, including (1) Registration Recall (*RR*), the percent of the successful registration satisfying the error thresholds of rotation and translation at the same time, (2) average Rotation Error (*RE*) and (3) average Translation Error (*TE*):

$$\text{RE}(\hat{\mathbf{R}}) = \arccos \frac{\text{Tr}\left(\hat{\mathbf{R}}^\top \mathbf{R}^*\right) - 1}{2}, \text{TE}(\hat{\mathbf{t}}) = \|\hat{\mathbf{t}} - \mathbf{t}^*\|_2^2,$$

(11)

where $\hat{\mathbf{R}}$ and $\hat{\mathbf{t}}$ are the predicted rotation matrix and rotation vector, respectively, while $\mathbf{R}^*$ and $\mathbf{t}^*$ are the corresponding ground truth. The average *RE* and *TE* are computed only on successful aligned point cloud pairs.

4.2. Comparison with Existing Methods

Evaluation on 3DMatch. We first evaluate our method on 3DMatch benchmark [50], which contains 46 training scenes, 8 validation scenes, and 8 test scenes. We first voxelize and down-sample the point cloud with 5cm voxel size and then leverage FCGF [11] and FPFH [37] descriptors to construct the putative correspondences based on the feature nearest neighbor. We compare our method with eight state-of-the-art (SOTA) correspondence-based methods, where FGR [52], SM [26], RANSAC (50k) [16], TEASER++ [45], and SC2_PCR [9] are representative traditional methods, while DGR [10], DHVR [25], and PointDSC [2] are advanced deep learning-based methods. As shown in Table 1,
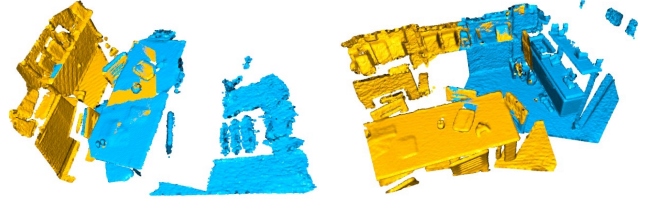

Figure 3. Registration visualization on 3DLoMatch [19].

in the FCGF setting, our method achieves the best performance in *RR* and *RE* criteria while the same *TE* with PointDSC. We need to highlight that *RR* is a more important criterion than *RE* and *TE* since the rotation and translation errors are just calculated in a successful registration. It means that the higher *RR* may include more challenging but successful registration cases, potentially increasing their errors. In the FPFH setting, it can be observed that our method can still achieve the best *RR* score among all deep methods, but perform slightly worse than SC2_PCR. Notably, compared to PointDSC (our baseline), the precision gain on *RR* is impressive (5.24%↑), which benefits from the effectiveness of our variational non-local feature learning and the voting-based inlier searching.

Evaluation on 3DLoMatch. We further test our method on 3DLoMatch benchmark [21]. Compared to 3DMatch sharing more than 30% overlap, the overlaps of point cloud pairs in 3DLoMatch just lie in $10\% \sim 30\%$, thus presenting much more challenges. We leverage FCGF [11] and recently popular Predator [21] as the feature descriptors for putative correspondence generation. We choose six traditional methods: FGR [52], SM [26], RANSAC (50k) [16], TEASER++ [45], SC2_PCR [9], and TR_DE [8], and two deep methods: DHVR [25], and PointDSC [2] for comparison. The registration recalls (*RR*) under different numbers of correspondences are listed in Table 2. It can be observed that regardless of FCGF or Predator descriptor, our method almost achieves the best performance on all settings, except for FCGF setting with 250 points. Notably, in more challenging cases, the performance advantage over PointDSC is further expanded (+9.5% and +7.8% in the cases of FCGF with 500 and 250 points). These experimental results further demonstrate the outstanding robustness of our method when encountering those extremely low-overlapping cases (high outlier ratios). Some visualization results are listed in Fig. 3 and the *RR* changes under different inlier ratios are presented in Fig. 4 (c), which suggests that our performance gains are mainly brought by model robustness in extremely high outlier situations.

Evaluation on KITTI. Finally, we evaluate our method on the outdoor LIDAR-scanned driving scenarios from KITTI dataset [19]. In line with [11], we utilize sequences 0-5, 6-7, and 8-10 as the training set, validation set, and test set, respectively. Also, as the setting in [3, 11], we further refine the ground-truth rigid transformations using ICP [5] and

Models	3DMatch (FCGF)			3DMatch (FPFH)			KITTI (FCGF)			KITTI (FPFH)			Sec.
	RR($\uparrow$)	RE($\downarrow$)	TE($\downarrow$)	RR($\uparrow$)	RE($\downarrow$)	TE($\downarrow$)	RR($\uparrow$)	RE($\downarrow$)	TE($\downarrow$)	RR($\uparrow$)	RE($\downarrow$)	TE($\downarrow$)	
FGR [52]	79.17	2.93	8.56	41.10	4.05	10.09	96.58	0.38	22.30	1.26	1.69	47.18	1.39
SM [26]	86.57	2.29	7.07	55.82	2.94	8.13	96.58	0.50	19.88	75.50	0.66	15.01	0.02
RANSAC [16]	91.50	2.49	7.54	73.57	3.55	10.04	97.66	**0.28**	22.61	89.37	1.22	25.88	6.43
TEASER++ [45]	85.77	2.73	8.66	75.48	2.48	7.31	83.24	0.84	**12.48**	64.14	1.04	14.85	0.07
DGR [10]	91.30	2.40	7.48	69.13	3.78	10.80	95.14	0.43	23.28	73.69	1.67	34.74	1.36
DHVR [25]	89.40	2.19	6.95	67.10	2.56	7.67	–	–	–	–	–	–	0.40
SC2_PCR [9]	<u>93.10</u>	**2.04**	<u>6.53</u>	**83.92**	<u>2.09</u>	<u>6.66</u>	<u>97.48</u>	0.33	20.66	97.84	<u>0.39</u>	9.09	0.09
PointDSC [2]	92.42	<u>2.05</u>	**6.49**	77.51	**2.08**	**6.51**	97.66	0.47	20.88	<u>98.20</u>	0.58	<u>7.27</u>	0.11
VBReg	**93.53**	**2.04**	**6.49**	<u>82.75</u>	2.14	6.77	**98.02**	<u>0.32</u>	20.91	**98.92**	0.32	**7.17**	0.22

Table 1. Quantitative comparison on 3DMatch [50] and KITTI [19] benchmark datasets with descriptors FCGF and FPFH. The registration speed is achieved by computing the averaged time cost on 3DMatch with FCGF descriptor.
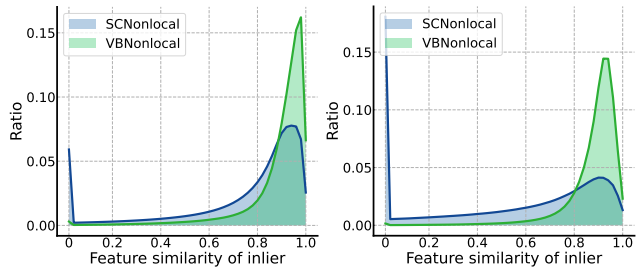
Feature	Model	5000	2500	1000	500	250
FCGF	FGR [52]	18.6	19.4	16.9	16.0	12.4
	SM [26]	32.4	31.3	31.4	28.0	23.5
	RANSAC [16]	37.6	37.2	35.9	32.1	25.9
	TEASER++ [45]	42.8	42.4	39.5	34.5	25.7
	DHVR [25]	50.4	49.6	46.4	41.0	<u>34.6</u>
	SC2_PCR [9]	<u>57.4</u>	<u>56.5</u>	<u>51.8</u>	<u>46.4</u>	**36.2**
	TR_DE [8]	49.5	50.4	48.4	43.4	34.3
	PointDSC [2]	55.8	52.6	46.8	37.7	26.7
	VBReg	**58.3**	**56.8**	**52.9**	**47.2**	34.5
Predator	FGR [52]	36.4	38.2	39.7	39.6	38.0
	SM [26]	53.8	55.1	55.4	54.5	50.2
	RANSAC [16]	62.3	62.8	62.4	61.5	58.2
	TEASER++ [45]	62.9	62.6	61.9	59.0	56.7
	DHVR [25]	67.2	67.3	66.1	64.6	60.5
	SC2_PCR [9]	<u>69.5</u>	<u>69.5</u>	<u>68.6</u>	<u>65.2</u>	<u>62.0</u>
	TR_DE [8]	64.0	64.8	61.7	58.8	56.5
	PointDSC [2]	68.1	67.3	66.5	63.4	60.5
	VBReg	**69.9**	**69.8**	**68.7**	**66.4**	**63.0**

Table 2. Registration recall (*RR*) with different numbers of points on 3DLoMatch benchmark dataset [21].

(a) The distribution of inlier feature similarity on 3DMatch [50].

(b) The distribution of inlier feature similarity on 3DLoMatch [21].

(c) *RR* under different inlier ratios.

(d) Training curves.

Figure 4. (a) and (b): The distribution of the learned feature similarity of inliers; (c): *RR* under different inlier ratios; (d): *RR* on the validation sets of 3DMatch [50] (FCGF) and 3DMatch (FPFH) under different training epochs.
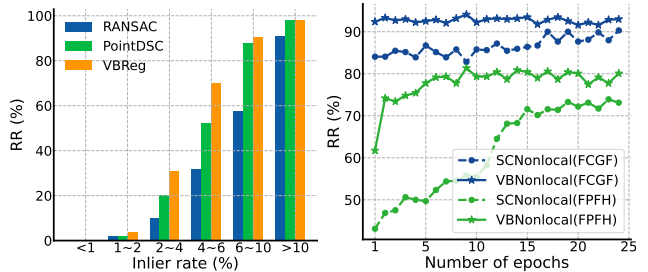
only collect the point cloud pairs far away from each other at most 10m as the test dataset. We downsample the point cloud with a voxel size of 30cm and exploit FCGF [11] and FPFH [37] descriptors for correspondence construction, respectively. The compared methods are consistent with those in *3DMatch*. The comparison results are listed in Table 1. For the FCGF setting, our method can achieve the best scores on the most important *RR* criterion while for the FPFH setting, our method can consistently achieve the best scores on all criteria.

4.3. Ablation Studies and Analysis

Variational Non-local Network. We first take PointDSC as our baseline and compare our proposed variational Bayesian non-local network (*VBNonlocal*) to the spatial consistency-based non-local network (*SCNonlocal*) to highlight the effectiveness of our proposed method. **(1)**: We

first compare their performance difference under two types of network input: *VBNonlocal^{xyz}* and *SCNonlocal^{xyz}* indicate using concatenated correspondence coordinate as input while *VBNonlocal^{feat}* and *SCNonlocal^{feat}* represent using concatenated coordinate and descriptor of correspondence as input. As shown in the top block in Table 3, with each data type as input, our *VBNonlocal* can consistently achieve significant performance gains. Especially, on 3DMatch with FPFH descriptor, *VBNonlocal^{feat}* brings 2.77% *RR* improvement and on 3DLoMatch with 500, 1000 and 2500 points, the *RR* improvements even can reach 5.4%, 6% and 5.9%, respectively. These impressive results support our view that our Bayesian-driven long-range dependency mod-

Model	3DMatch		3DLoMatch (FCGF)					3DLoMatch (Predator)					Sec.
	FCGF	FPFH	5000	2500	1000	500	250	5000	2500	1000	500	250	
PointDSC *w/* SCNonlocal^{xyz}	92.42	77.51	55.8	52.6	46.8	37.7	26.7	68.1	67.3	66.5	63.4	60.5	0.11
PointDSC *w/* VBNonlocal^{xyz}	93.04	80.16	57.7	55.6	50.2	39.9	26.1	69.7	69.6	67.9	64.9	61.9	0.17
PointDSC *w/* SCNonlocal^{feat}	92.36	77.76	54.6	50.6	44.9	36.8	25.4	69.2	68.6	67.9	63.5	59.9	0.13
PointDSC *w/* VBNonlocal^{feat}	93.04	80.53	56.9	56.5	50.9	42.2	28.9	69.2	68.7	68.0	64.6	60.6	0.18
PointDSC *w/* SCNonlocal^{cls}	92.98	78.99	54.1	52.2	46.0	38.7	27.7	67.6	66.9	67.2	63.7	60.2	0.11
PointDSC *w/* VBNonlocal^{feat}+Vote	93.41	81.21	58.3	56.5	51.9	44.7	31.1	69.3	69.5	68.2	65.3	61.2	0.20
PointDSC *w/* VBNonlocal^{feat}+Vote+CS	**93.53**	**82.75**	**58.3**	56.8	52.9	47.2	34.5	**69.9**	69.8	68.7	66.4	63.0	0.22
Iteration times $L = 6$	93.41	82.32	58.1	**57.1**	52.9	48.3	34.8	69.7	69.7	68.7	66.3	63.8	0.19
Iteration times $L = 9$	93.41	81.58	58.2	57.0	53.2	47.6	32.5	**70.0**	69.4	68.5	66.9	63.2	0.20
Iteration times $L = 12$*	**93.53**	**82.75**	**58.3**	56.8	52.9	47.2	34.5	69.9	69.8	68.7	66.4	63.0	0.22
Random feat. dim. $\tilde{d} = 32$	93.41	82.38	58.0	**57.5**	53.5	48.1	34.8	69.7	70.0	68.6	66.3	63.3	0.20
Random feat. dim. $\tilde{d} = 64$	93.41	81.45	57.9	56.9	52.8	**48.6**	35.0	69.6	69.7	68.4	66.5	62.3	0.21
Random feat. dim. $\tilde{d} = 128$*	**93.53**	**82.75**	**58.3**	56.8	52.9	47.2	34.5	**69.9**	69.8	69.3	66.4	63.3	0.22

Table 3. Ablation studies on 3DMatch [50] and 3DLoMatch [21] datasets. *SCNonlocal*: Spatial consistency-guided non-local network; *VBNonlocal*: Variational Bayesian-based non-local network; *Vote*: Voting-based inlier searching; *CS*: Conservative seed selection.

eling can effectively learn the discriminative inlier/outlier features for reliable inlier search. **(2)**: Then, to further highlight the superiority of our variational inference-guided feature learning, we also try to add classification loss on the features produced by each iteration in *SCNonlocal* to guide their learning (denoted as *SCNonlocal^{cls}*). As shown in the fifth row in Table 3, such loss-based label-propagation way just can achieve very limited performance gain and even sometimes degrades score. It demonstrates the superiority of our label-dependent posterior guidance for (prior) feature learning. Also, owing to such posterior guidance, the training curves in Fig. 4 (d) show that our method can achieve significantly faster convergence speed than *SCNonlocal*.

Discriminative Feature Learning? In order to verify whether *VBNonlocal* can learn more discriminative features than *SCNonlocal*, we visualize the distribution of the feature similarity of inliers on 3DMatch (Fig. 4 (a)) and 3DLoMatch (Fig. 4 (b)). As we can see, our learned inlier features own much higher similarities (approximate to 1) than *SCNonlocal* on both datasets, which demonstrates that our proposed Bayesian-inspired non-local network truly can promote more discriminative correspondence-feature learning.

Variational Non-local Setting. We further test the performance changes under different *VBNonlocal* settings. **(1)** We first test the model robustness under different numbers of non-local iterations. The second block in Table 3 verifies that our method is robust to the iteration time and tends to consistently achieve outstanding *RR* score. **(2)** Then, the bottom block in Table 3 further shows our model stability to different dimensions of random features.

Voting-based Inlier Searching. Furthermore, we evaluate the performance contribution of the proposed voting-based inlier searching strategy (*Vote*). As we can see in the sixth row of Table 3, voting strategy can consistently achieve performance improvement regardless in 3DMatch or in more challenging 3DLoMatch. It mainly benefits from the high-quality *hypothetical inliers* sampled by our voting policy.

Conservative Seed Selection. Finally, we test the effectiveness of the proposed conservative seed selection strategy (*CS*) motivated by Theorem 1. As we can see in the seventh row of Table 3, *CS* can achieve consistent performance gain in each setting. Especially, in the cases of fewer points (*e.g.*, FCGF setting with 250 and 500 points), the improvement is much more significant (+3.4%, +2.5%). As the analysis before, in sparse point clouds, the original inlier selection strategy like in [2] is aggressive and prone to miss too many inlier seeds. Instead, our conservative selection strategy can effectively mitigate it as well as keep registration efficiency.

5. Conclusion

In this paper, we adapted the variational Bayesian inference into the non-local network and developed the effective Bayesian-guided long-term dependencies for discriminative correspondence-feature learning. To achieve effective variational inference, a probabilistic graphical model was customized over our non-local network, and the variational low bound was derived as the optimization objective for model training. In addition, we proposed a Wilson score-based voting mechanism for high-quality inlier sampling and theoretically verified its superiority over RANSAC. Extensive experiments on indoor/outdoor datasets demonstrated its promising performance.

6. Acknowledgments

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM (2011)*, 54(10):105–112. 1

[2] Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. Pointdsc: Robust point cloud registration using deep spatial consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15859–15869, 2021. 1, 2, 3, 5, 6, 7, 8

[3] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3feat: Joint learning of dense detection and description of 3d local features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6359–6367, 2020. 1, 2, 6

[4] Daniel Barath and Jiří Matas. Graph-cut ransac. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6733–6741, 2018. 2, 6

[5] Paul J Besl and Neil D McKay. Method for registration of 3-D shapes. In *Sensor fusion IV: control paradigms and data structures (1992)*, volume 1611, pages 586–606. 6

[6] Eric Brachmann and Carsten Rother. Neural-guided ransac: Learning where to sample model hypotheses. In *ICCV*, 2019. 2

[7] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000. 1

[8] Wen Chen, Haoang Li, Qiang Nie, and Yun-Hui Liu. Deterministic point cloud registration via novel transformation decomposition. In *CVPR*, 2022. 6, 7

[9] Zhi Chen, Kun Sun, Fan Yang, and Wenbing Tao. Sc2-pcr: A second order spatial compatibility for efficient and robust point cloud registration. In *CVPR*, 2022. 2, 6, 7

[10] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *CVPR (2020)*, pages 2514–2523. 1, 2, 6, 7

[11] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8958–8966, 2019. 1, 2, 6, 7

[12] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 4

[13] Zheng Dang, Lizhou Wang, Yu Guo, and Mathieu Salzmann. Learning-based point cloud registration for 6d object pose estimation in the real world. In *ECCV*, 2022. 1, 2

[14] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 602–618, 2018. 1

[15] Jean-Emmanuel Deschaud. IMLS-SLAM: Scan-to-model matching based on 3D data. In *ICRA (2018)*. 1

[16] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM (1981)*, 24(6):381–395. 1, 2, 5, 6, 7

[17] Andrea Frome, Daniel Huber, Ravi Kolluri, Thomas Bülow, and Jitendra Malik. Recognizing objects in range data using regional point descriptors. In *European conference on computer vision*, pages 224–237. Springer, 2004. 1, 2

[18] Kexue Fu, Shaolei Liu, Xiaoyuan Luo, and Manning Wang. Robust point cloud registration framework based on deep graph matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8893–8902, 2021. 2

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 6, 7

[20] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 5

[21] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4267–4276, 2021. 1, 2, 3, 6, 7, 8

[22] Haobo Jiang, Jianjun Qian, Jin Xie, and Jian Yang. Planning with learned dynamic model for unsupervised point cloud registration. *IJCAI (2021)*. 2

[23] Haobo Jiang, Yaqi Shen, Jin Xie, Jun Li, Jianjun Qian, and Jian Yang. Sampling network guided cross-entropy method for unsupervised point cloud registration, 2021. 2, 6

[24] Huu M Le, Thanh-Toan Do, Tuan Hoang, and Ngai-Man Cheung. SDRSAC: Semidefinite-based randomized approach for robust point cloud registration without correspondences. In *CVPR (2019)*. 2

[25] Junha Lee, Seungwook Kim, Minsu Cho, and Jaesik Park. Deep hough voting for robust global registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15994–16003, 2021. 6, 7

[26] Marius Leordeanu and Martial Hebert. A spectral technique for correspondence problems using pairwise constraints. 2005. 6, 7

[27] Jiayuan Li, Qingwu Hu, and Mingyao Ai. Gesac: Robust graph enhanced sample consensus for point cloud registration. *ISPRS Journal of Photogrammetry and Remote Sensing*, 167:363–374, 2020. 2

[28] Jiahao Li, Changhao Zhang, Ziyao Xu, Hangning Zhou, and Chi Zhang. Iterative distance-aware similarity matrix convolution with mutual-supervised point elimination for efficient point cloud registration. In *ECCV (2020)*. 2

[29] Xiang Li, Lingjing Wang, and Yi Fang. PC-Net: Unsupervised point correspondence learning with neural networks. In *3DV (2019)*. 2

[30] Xiang Li, Lingjing Wang, and Yi Fang. Unsupervised partial point set registration via joint shape completion and registration. *arXiv preprint arXiv:2009.05290 (2020)*. 2

[31] Yang Li and Tatsuya Harada. Lepard: Learning partial point cloud matching in rigid and deformable scenes. In *CVPR*, 2022. 2

[32] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 1

[33] G Dias Pais, Srikumar Ramalingam, Venu Madhav Govindu, Jacinto C Nascimento, Rama Chellappa, and Pedro Miraldo. 3DRegNet: A deep neural network for 3D point registration. In *CVPR (2020)*. 1, 2

[34] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR (2017)*, pages 652–660. 2

[35] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *CVPR*, 2022. 2

[36] Siwen Quan and Jiaqi Yang. Compatibility-guided sampling consensus for 3-d point cloud registration. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 1

[37] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE international conference on robotics and automation*, 2009. 1, 2, 6, 7

[38] Samuele Salti, Federico Tombari, and Luigi Di Stefano. Shot: Unique signatures of histograms for surface and texture description. *Computer Vision and Image Understanding*, 125:251–264, 2014. 1, 2

[39] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR (2016)*. 1

[40] Yaqi Shen, Le Hui, Haobo Jiang, Jin Xie, and Jian Yang. Reliable inlier evaluation for unsupervised point cloud registration. *arXiv preprint arXiv:2202.11292*, 2022. 2

[41] Haiping Wang, Yuan Liu, Zhen Dong, Wenping Wang, and Bisheng Yang. You only hypothesize once: Point cloud registration with rotation-equivariant descriptors. *arXiv preprint arXiv:2109.00182*, 2021. 1, 2

[42] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *ICCV (2019)*. 2

[43] Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 1927. 5

[44] Jay M Wong, Vincent Kee, Tiffany Le, Syler Wagner, Gian-Luca Mariottini, Abraham Schneider, Lei Hamilton, Rahul Chipalkatty, Mitchell Hebert, David MS Johnson, et al. Segicp: Integrated deep semantic segmentation and pose estimation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017. 1

[45] Heng Yang, Jingnan Shi, and Luca Carlone. Teaser: Fast and certifiable point cloud registration. *IEEE Transactions on Robotics*, 2020. 2, 6, 7

[46] Jiaqi Yang, Ke Xian, Peng Wang, and Yanning Zhang. A performance evaluation of correspondence grouping methods for 3d rigid data matching. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1

[47] Zi Jian Yew and Gim Hee Lee. RPM-Net: Robust point matching using learned features. In *CVPR (2020)*. 2

[48] Zi Jian Yew and Gim Hee Lee. Regtr: End-to-end point cloud correspondences with transformers. In *CVPR*, 2022. 2

[49] Kwang Moo Yi, Eduard Trulls, Yuki Ono, Vincent Lepetit, Mathieu Salzmann, and Pascal Fua. Learning to find good correspondences. In *CVPR*, 2018. 2

[50] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1802–1811, 2017. 1, 2, 6, 7, 8

[51] Ji Zhang and Sanjiv Singh. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and Systems*, 2014. 1

[52] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *ECCV*, 2016. 2, 6, 7

[53] Jing Zhu and Yi Fang. Reference grid-assisted network for 3D point signature learning from point clouds. In *WACV (2020)*. 2